AFRL-RI-RS-TR-2018-030

DIVERSE DATA SETS FOR IMPACT

UNIVERSITY OF WISCONSIN SYSTEM

FEBRUARY 2018

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

AFRL-RI-RS-TR-2018-030 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /	**/ S /**
ROBERT L.KAMINSKI	WARREN H. DEBANY JR
Work Unit Manager	Technical Advisor, Information Exploitation and Operations Division Information Directorate

REPORT DOCUMENTATION PAGE	*Form Approved* **OMB No. 0704-0188**

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
FEB 2018	FINAL TECHNICAL REPORT	SEP 2012 – SEP 2017

4. TITLE AND SUBTITLE

DIVERSE DATA SETS FOR IMPACT

5a. CONTRACT NUMBER
N/A

5b. GRANT NUMBER
FA87A50-12-2-0328

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Paul Barford

5d. PROJECT NUMBER
HS53

5e. TASK NUMBER
WI

5f. WORK UNIT NUMBER
SC

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Wisconsin System
Suite 6401
21 North Park Street
Madison WI 53715-1218

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RIG
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)

AFRL/RI

11. SPONSOR/MONITOR'S REPORT NUMBER

AFRL-RI-RS-TR-2018-030

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The efforts on this grant have focused on data distribution and research. The developed Internet Atlas portal received over 20K page views by over 11K unique visitors, Effort provided over 50 accounts for access to Internet Atlas, and from Jan. '16 – May '17, 119 datasets were provided (primarily DSHIELD logs and the Internet Long Haul infrastructure data). Effort had 14 research publications in high quality venues. These research activities have focused on understanding details of the Internet's physical infrastructure; understanding the perspective on Internet events provided by NTP measurements and developing methods to improve Internet time synchronization; and on using crawlers to reveal details of aspects of web organization and use. Internet Atlas itself has been the subject of numerous articles in both technical and popular press, and has resulted in awards such as Popular Science's Greatest Innovations of 2017. Atlas has also been presented in the RSA conference, the DHS showcase and is being considered for commercialization.

15. SUBJECT TERMS
Internet measurement, Internet physical layer topology, Internet maps, Internet risk analysis, Network Time Protocol, Internet outage analysis, Web crawling

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					ROBERT L. KAMINSKI
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**	SAR	15	**19b. TELEPHONE NUMBER (*Include area code)***
U	U	U			

Table of Contents

Section **Page**

Contents

1 SUMMARY

Over the past four years, the University of Wisconsin has been a performer in the DHS IMPACT (formerly PREDICT) program, which seeks to provide data sets to the community for cyber security research. The primary focus of our activities has been development of a novel repository of maps of physical Internet infrastructure and a web portal for visualization and analysis of the maps. The combination of the repository and portal is called *Internet Atlas*. In addition to Internet Atlas, the University of Wisconsin has developed and distributed several other data sets including DSHIELD IDS/firewall logs, BGP updates, Network Time Server logs and web crawl logs.

We have distributed hundreds of data sets to the research community over the past four years. For the period from May '13 through May '17, the Internet Atlas portal received over 20K page views by over 11K unique visitors from all over the world. Over 50 accounts for detailed access to Atlas have been provided. For the period from Jan. '16 – May '17 (during which time distribution was tracked in detail), 119 datasets were provided (primarily DSHIELD logs and the Internet Long Haul infrastructure data).

In addition to distributing data sets, Atlas data and other data assembled for this project has been the basis for 14 research publications in high quality venues (two additional manuscripts are under submission) from our group. A complete list of publications that resulted from this grant are provided in Section 6. These research activities and Internet Atlas itself have been the subject of numerous articles in both technical and popular press (*e.g.,* articles in the NY Times, Boston Globe, Tech Republic, VoA, Business Insider, Gizmodo, Mashable, etc.). Atlas has also been presented in the RSA conference, the DHS showcase and is being considered for commercialization.

2 INTRODUCTION

Over the past four years and through support from the DHS IMPACT program, the University of Wisconsin has developed and distributed unique data sets to the research community. The primary focus of our efforts has been on development *Internet Atlas*, which includes a novel repository of maps of *physical Internet infrastructure* (defined as fiber conduits and buildings where conduits terminate) and a web portal for visualization and analysis of the maps and for real-time active measurement. The distinguishing features of Internet Atlas are its scale (it's the largest repository of Internet maps in the world including over 1.4K maps of individual networks), the variety of data in the repository (BGP, NTP, etc.) and that all data is normalized and displayed in a consistent, geocoded format.

The starting point for Atlas is a geographically anchored representation of the physical Internet including *(i)* nodes (*e.g.,* co-location centers, hosting facilities and data centers), *(ii)* fiber conduits/links that connect these nodes, and *(iii)* relevant meta data (*e.g.,* source provenance). This physical representation is built by using search to identify primary source data such as maps and other repositories of service provider network information. This data is then carefully entered into the database using a combination of manual and automated processes including consistency checks and methods for geocoding both node and link data. Customized interfaces were built to import a variety of dynamic (*e.g.*, BGP updates, weather updates, etc.) and static (*e.g.*, highway, rail, census, etc.) data into Atlas, and to layer it on top of the physical representation. The openly available web portal (www.internetatlas.org) is based on the widely-used ArcGIS geographic information system, which enables visualization and diverse spatial analyses of the data. Network maps and other relevant data (*e.g.*, cell towers, public WiFi locations and other physical infrastructure) continue to be added to the repository on an on-going basis. New measurement and analysis capabilities (*e.g.*, network monitoring via NTP) will be added as well.

Distribution of data from this project has focused on Internet Atlas and the other data sets mentioned above. The primary interest in terms of requests has been on the map and associated data related to the US Long Haul Fiber Infrastructure identified in our ACM SIGCOMM '15 paper. That map identifies for the first time, the main data carrying component of the Internet in the US. We have also provided accounts on the Atlas portal to over 50 requesters over the performance period. The portal received regular, daily use. We have also provided the DSHIELD intrusion detection and firewall logs many times over the performance period. The remaining data sets provided by the University of Wisconsin have only been requested a number of times.

Research conducted during the period of performance has focused on aspects of the Internet's physical-layer topology, the Network Time Protocol (NTP) and the web. Our efforts on understanding physical-layer connectivity have been realized in Internet Atlas, but have also included studies that clarify outage risk due to natural disaster events, active probe-based measurement capability that can expand and enhance Internet maps, methods for identifying where new connectivity can be deployed to improve robustness and performance, and methods for identifying where new connectivity can be deployed in underserved areas. Our efforts on the NTP recognize that this protocol is unique in that it is one of the few on-by-default protocols in the Internet and therefore offers a unique opportunity for opportunistic measurement. Our studies have revealed the ability of NTP data to provide high-accuracy latency measurements at scale and without the need for deployment of any specialized infrastructure. We have also identified how NTP can be enhanced to support high quality timing in mobile and wireless environments. Our latest work on NTP shows how it can be used to identify anomalies and outages throughout the Internet. Finally, our work on the web has focused on using crawling-based measurement to reveal details of web characteristics including the Internet Adscape i.e., characteristics of ads delivered to users, and how third party tracking cookies reveal user characteristics without disclosure to users.

Our research efforts have resulted 14 published research papers that have appeared in prestigious venues such as the ACM SIGCOMM Conference, ACM Internet Measurement Conference, the ACM HotNets Conference, World Wide Web Conference, the AAAI Conference and others. Details are listed in Section 7. These paper include one best paper award (ACM SIGCOMM GAIA '16 Workshop), one Best of What's New award (Popular Science '17) and numerous mentions in major publications. The work on Atlas has also been the subject of numerous presentations *e.g.*, at the RSA conference, the DHS showcase and at a number of companies and universities.

3 METHODS

Building Internet Atlas began by using Internet search engines to find any and all maps/listings of Internet infrastructure. Several important lessons were learned through extensive trial-and-error exploration of relevant search terms. For example, simple one-word terms, such as “co-location” or “datacenter”, resulted in discovery of very few previously unseen networks/locations, while multiple word phrases, such as “co-location facility” or “telecom hotels” were more productive. The most important lesson learned is that geographic specificity in search terms is extremely important in revealing regional and local providers. While this may seem obvious, it is complicated by the vast number of local service providers that are only concerned with last mile connectivity.

In addition to Internet search, we appealed to the large number of existing Internet systems and publicly available data they provide. This includes Peering DB, Network Time Protocol (NTP) servers, Domain Name System servers (DNS), listings of Internet Exchange Points (IXPs), Looking Glass servers, traceroute servers, Network Access Points, etc. Beyond their intrinsic interest, it is important to recognize that NTP servers often publish their Lat/Lon coordinates and are typically co-located with other networking/computing equipment. Similarly, DNS servers routinely publish their location via the LOC record. In total, over 4, 700 network resources of various types are annotated in the Internet Atlas database.

Once a target network has been discovered via search, we transcribe the information to Atlas’ GIS database. This is complicated by the varying data formats used by each provider. Network maps can range from images (such as the Sprint Network Map), to interactive maps (such as, the Flash-based AT&T Map and the Google Maps-based Level3 Map).

Visualization-centric representations often reveal no information about link paths other than connectivity (*e.g.,* line-of-sight abstractions are common). For these we enter the network adjacency graph by hand into Atlas. However, some maps provide highly detailed geographic layouts of fiber conduit connectivity (*e.g.,* Level3). We transcribe these, maintaining geographic accuracy, into the Atlas using a process and scripts that *(i)* capture high resolution sub-images, *(ii)* patch sub-images into a composite image, *(iii)* extract a network link image using color masking techniques, *(iv)* project the link-only image into ArcGIS using geographic reference points (*e.g.,* cities), and *(v)* use link vectorization in ArcGIS to enable analysis (*e.g.,* distance estimation) of the links.

Node locations in primary source data are provided in four forms: Lat/Lon, street address, city or state. If none of these location types is provided, then the node is not entered into the repository. All node locations are geo-coded into the Atlas repository as a Lat/Lon, while maintaining the source information as metadata. If a Lat/Lon for a network resource is provided, that is transcribed directly into the repository. If a street address for a resource is provided, that address is translated into a Lat/Lon using ArcGIS’ inherent capabilities. If only a city/state location is provided, then that is translated into the Lat/Lon of the city/state center if no other more specific addresses for network infrastructure are available in that city/state. Otherwise, the Lat/Lon of the location in that city/state that has the most references from other networks is used. While clearly this could be inaccurate, we believe it is likely to be more accurate than simply leaving the location in the city/state center.

Provider maps often contain additional information about network node resources. This information can range from location (potentially down to Lat/Lon coordinates), to IP addresses, to resource or service types. Our ability to extract network node information from the discovered resources is dependent on an assembly of scripts that include Flash-based extraction and parsing tools, optical character recognition parsing tools, PDF-based parsing tools, in addition to standard text manipulation tools. This library of parsing scripts can extract information and enter it into database automatically. For instances where none of the tools or scripts are successful on the provider data, we manually parse and enter the data.

Despite our efforts to automate the data transcription process, we still have to enter data manually from time to time. Thus, we must account for human error in this process. We employ several completeness and consistency checks to verify that that manually entered data accurately reflects the primary source data. These checks include having someone other than the person who entered the data manually compare it to the source data. We also conducted a complete audit of the Atlas data repository in 2016. This process included appealing to the most recent version of the primary source data to ensure that the representation in Atlas was accurate. While this process resulted in very few changes of the node/link data, over 120 networks in the repository had undergone some kind of management change (typically an acquisition by a larger entity) since original entry.

While the link location data gathered from search are usually reliable due to the stability and static nature of the underlying fiber infrastructure, we collected additional information sources to validate these data. We also use these additional information sources to infer whether some links follow the same physical ROW, which indicates that the fiber links either reside in the same fiber bundle, or in an adjacent conduit. In this case, we use a variety of public records to geolocate and validate link endpoints and conduits. These records tend to be rich with detail and have not been discussed in prior work. In addition to maps, we have found a large number of these documents, which we have used to validate over 90% of our long haul infrastructure data.

ASSUMPTIONS

Internet Atlas is predicated on the assumption that detailed information on network infrastructure can be found on publicly available webpages. This implies that Internet search can be used as the primary data gathering tool. In addition to the major search engines, we used search aggregators, such as Soovle and SidePad to enhance the ability to find network maps.

Our working assumption on sources of data that we use to validate and identify link/conduit locations is that ISPs, government agencies, and other relevant parties often archive documents on public-facing websites. Specifically, we seek information that can be extracted from government agency filings, environmental impact statements, documentation released by third-party fiber services, indefeasible rights of use (IRU) agreements, press releases, and other related resources.

PROCEDURES

Assembly of data in the Internet Atlas repository is the starting point for much of the research and data distribution by the University of Wisconsin over the period of performance. Beyond Atlas, other data sets that have been assembled and distributed relate to research activities on the *(i)* Network Time Protocol (NTP) logs provider by NTP server operators and *(ii)* web cookies collected by crawling the Alexa top 100K sites.

Data related to Atlas was distributed in two ways. First, password protected access to the portal is granted on request for the purpose of research. Second, the long haul fiber infrastructure data was assembled into a separate distribution which was provided to researchers on request. The NTP and web cookie data is also package into compressed distributions, which have been provided on request via the IMPACT portal. Research over the period of performance was conducted in three general area related to each of the three aforementioned data sets: physical-layer connectivity based on Atlas data, network timing and latency based on NTP and web behavior based on crawling sites.

Research procedures on the physical layer have focused on *(i)* understanding the details of the characteristics of connectivity, *(ii)* understanding risks, *(iii)* identifying opportunities to improve robustness and performance. In each case, our procedures begin with Internet map data from Atlas and focus on developing novel methods for analysis. For example, we our high-level definition of a long-haul link4 is one that connects major city-pairs. In order to be consistent when processing existing map data, however, we used the following concrete definition. We define a long-haul link as one that spans at least 30 miles, or that connects population centers of at least 100,000 people, or that is shared by at least 2 providers. Using this definition, we were able to extract long haul fiber infrastructure from a subset of maps in Atlas. In a number of our studies, we have utilized third party data collected at higher layers of the protocol stack (e.g., CAIDA's Ark data or active probe-based measurements or BGPmon data) that we connect to physical layer representations and thereby reveal key characteristics.

Research procedures on NTP have focused on *(i)* improve time synchronization in the Internet and *(ii)* utilizing measurements of NTP to enhance our understanding of the Internet. Our research efforts recognize the NTP is an extremely important protocol – providing time synchronization for all connected devices – and that its on-by-default nature makes it a unique source of opportunistic measurement. The procedures that we created for enhancing timing have focused on wireless and IoT devices, which typically have lower quality crystals and therefore are more challenging to synchronize. In each case, we conducted laboratory-based experiments on novel protocols using insights derived from our empirical study of NTP. Our research on utilizing NTP to provide insights on the Internet have focused on the fact that embedded in NTP packets are details on one-way-delays between hosts that can provide a unique and valuable perspective on Internet latency. Our studies have focused on utilizing this data to provide insights on path properties, Internet coordinate systems and Internet outages. In each case, we have developed novel methods to extract key details from NTP packet exchanges.

Research procedures on the web have been relatively limited and have focused primarily on utilizing data gathered through large-scale crawling campaigns. Web site crawling remains one of the best ways to gather data on web characteristics and to develop an understanding of how to improve this infrastructure. Our procedures have focused on creating capabilities to extract key information such as where and how ads appear on web sites. These procedures include the ability to build profile-aware crawlers that result in different ads and content being delivered depending on the profile. Our procedures also focus on how trackers that are commonly deployed on web sites. Our analysis procedures have focused on analyzing how these trackers used by third parties can expose user characteristics and thereby compromise privacy.

4 RESULTS AND DISCUSSION

Key results from our studies of physical layer Internet topology include the following:

1) We find that there are substantially more nodes and links identified in the service provider map data versus the active probe data gathered at layer 3 (e.g., CAIDA's Ark project). We developed a new method for probe-based measurement of physical infrastructure called POPsicle that is based on careful selection of probe source-destination pairs. We demonstrated the capability of our method through an extensive measurement study using existing "looking glass" vantage points distributed throughout the Internet and show that it reveals 2.4 times more physical node locations versus standard active probing methods. To demonstrate the deployability of POPsicle we also conducted tests at an IXP. Our results again show that POPsicle can identify more physical node locations compared with standard layer 3 probes, and through this deployment approach it can be used to measure thousands of networks world-wide.

2) Our study on the Internet's long-haul fiber-optic infrastructure in the US resulted in building a first-of-its-kind map. We validated the map rigorously by appealing to public information sources such as government agency filings, environmental impact statements, press releases, and others. Examination of the map confirmed the close correspondence of fiber deployments and road/rail infrastructure and reveals significant link sharing among providers. We also applied different metrics to examine the issue of shared risk in the long-haul map. Our results point to high-risk links where there are significant levels of sharing among service providers. Finally, we identified public ROWs that could be targets for new link conduits that would reduce shared risk and improve path performance. We also point out how our findings expand the current discussion on how Title II and net neutrality.

3) We developed a framework for identifying target areas for new network infrastructure deployment in underserved areas. Our approach considers *(i)* infrastructure availability, *(ii)* user demographics, and *(iii)* deployment costs. We use multi-objective optimization to identify geographic areas that have the highest concentrations of un/underserved users and that can be upgraded at the lowest cost. We demonstrated the efficacy of our framework by consider physical infrastructure and demographic data from the US and two different deployment cost models. Our results identify a list of counties that would be attractive targets for broadband deployment from both cost and impact perspectives. We compare the areas that our framework identify with those identified in the FCC's

broadband report and find strong overlap between the two highlighting the opportunity to apply our method in future studies.

Key results from our studies of the Network Time Protocol include the following:

1) We investigated the efficacy of using a novel but non-obvious source of data for studying Internet latency—logs from network time protocol (NTP) servers. We conducted this study by analyzing logs collected from 10 NTP servers distributed across the United States. These logs include over 73M latency measurements to 7.4M worldwide clients (as indicated by unique IP addresses) collected over the period of one day. Our initial analysis of the general characteristics of propagation delays derived from the log data reveals that delay measurements from NTP must be carefully *filtered* in order to extract accurate results. We developed a filtering process that removes measurements that are likely to be inaccurate. After applying our filter to NTP measurements, our analysis reveals a wide range of observed latencies across all servers. For example, around 99% of US-based clients have latencies less than 100 milliseconds to the server with which they synchronize, compared to an earlier survey from 1999 which showed that 90% of clients had latencies below 100 milliseconds to their server. We also observe a highly diverse client-base from a geographic perspective, especially for the secondary (stratum-2) servers. This diversity is much less pronounced for primary (stratum-1) servers because they are more tightly controlled, and for IPv6-based servers since they generally have a smaller set of clients they serve.

2) We studied clock synchronization in mobile hosts, which often implement a simplified version of the Network Time Protocol (NTP), known as SNTP, due to resource constraints typical of mobile devices. We reported an analysis of logs from NTP servers that highlights the significant differences in synchronization behavior of wireline vs. wireless hosts. This analysis motivated our laboratory-based study of the details of clock synchronization on mobile hosts, which revealed the causes and extent to which synchronization can become misaligned. We then developed a new protocol that we call Mobile NTP (MNTP), which is designed to be simple, efficient and easy to deploy. We implemented MNTP on a wireless laptop and demonstrated its capability over a range of operating conditions. We found that MNTP maintains clock synchronization to within 25ms of a reference clock, which is over 12 times better than standard SNTP.

3) We developed a new approach to Internet event measurement, identification and analysis that provides a broad, detailed and accurate perspective without the need for new or dedicated infrastructure or additional network traffic. Our approach is based on analyzing data that is readily available from NTP servers. We develop a tool for analyzing NTP traces, which is based on Robust Principal Components Analysis. We demonstrated the utility of this tool by applying it to data collected over a period of 3 months at 19 NTP servers. We reported on the characteristics of the events identified in our analysis. We also compared and contrasted the perspective provided by our tool with events identified by standard ping-based active probing, reported outages, and a simple low-pass filter style event detector. We found that while there is commonality across

methods, NTP-based monitoring provides a perspective that is unique and complements prior methods.

Key results from our studies of the web include the following:

1) We conducted a first-of-its-kind study to broadly understand the features, mechanisms and dynamics of display advertising on the web - *i.e.*, the *Adscape*. Our study took the perspective of users who are the targets of display ads shown on web sites. We developed a scalable crawling capability that enables us to gather the details of display ads including creatives and landing pages. Our crawling strategy focused on maximizing the number of unique ads harvested. Of critical importance was the recognition that a user's profile (*i.e.*, browser profile and cookies) can have a significant impact on which ads are shown. We deployed our crawler over a variety of websites and profiles and this yields over 175K distinct display ads. We found that while targeting is widely used, there remain many instances in which delivered ads do not depend on user profile; further, ads vary more over user profiles than over websites. We also assessed the population of advertisers seen and identify over 3.7K distinct entities from a variety of business segments. Finally, we found that when targeting is used, the specific types of ads delivered generally correspond with the details of user profiles, and also on users' patterns of visit.

2) We conducted an empirical study of web cookie characteristics, placement practices and information transmission. To conduct this study, we implemented a lightweight web crawler that tracks and stores the cookies as it navigates to websites. We used this crawler to collect over 3.2M cookies from the two crawls, separated by 18 months, of the top 100K Alexa web sites. We analyzed the general cookie characteristics and added context via a cookie category index and website genre labels. We considered privacy implications by examining specific cookie attributes and placement behavior of 3rd party cookies. We found that 3rd party cookies outnumber 1st party cookies by a factor of two, and we illuminate the connection between domain genres and cookie attributes. We found that less than 1% of the entities that place cookies can aggregate information across 75% of web sites. Finally, we considered the issue of information transmission and aggregation by domains via 3rd party cookies. We developed a mathematical framework to quantify user information leakage for a broad class of users. In particular, we demonstrated the interplay between a domain's footprint across the Internet and the browsing behavior of users has significant impact on information transmission.

5 CONCLUSION

Over the course of the past four years, our efforts on this grant have focused on data distribution and research We have distributed hundreds of data sets to the research community. For the period of performance, the Internet Atlas portal received over 20K page views by over 11K unique visitors, we provided over 50 accounts for access to Internet Atlas, and from Jan. '16 – May '17 (during which time distribution was tracked in detail), 119 datasets were provided (primarily DSHIELD logs and the Internet Long Haul infrastructure data).

In addition to distributing data sets, we have had 14 research publications in high quality venues (two additional manuscripts are under submission). These research activities have focused on understanding details of the Internet's physical infrastructure; understanding the perspective on Internet events provided by NTP measurements and developing methods to improve Internet time synchronization; and on using crawlers to reveal details of aspects of web organization and use. Internet Atlas itself has been the subject of numerous articles in both technical and popular press, and has resulted in awards such as Popular Science's Greatest Innovations of 2017. Atlas has also been presented in the RSA conference, the DHS showcase and is being considered for commercialization.

6 REFERENCES

1) Ram Durairajan, Subadip Ghosh, Xin Tang, Paul Barford and Brian Eriksson "Internet Atlas: A Geographic Database of the Internet", In Proceedings of the 5th ACM SIGCOMM HotPlant Workshop Conference, Hong Kong, August, 2013.

2) Brian Eriksson, Ram Durairajan and Paul Barford. "RiskRoute: A Framework for Mitigating Network Outage Threats", In Proceedings of ACM CoNEXT, Santa Barbara, CA, December, 2013.

3) Ram Durairajan, Joel Sommers and Paul Barford. "Layer 1-Informed Internet Topology Measurement", In Proceedings of the ACM Internet Measurement Conference, Vancouver, BC, November, 2014.

4) Ram Durairajan, Paul Barford, Joel Sommers and Walter Willinger. "InterTubes: A Study of the US Long-haul Fiber-optic Infrastructure", In Proceedings of ACM SIGCOMM, London, UK, August 2015.

5) Ram Durairajan, Sathiya Kumaran, Joel Sommers and Paul Barford. "Times Forgotten: Using NTP to understand Internet Latency", In Proceedings of ACM Workshop on Hot Topics in Networks (HotNets), Philadelphia, PA, November, 2015.

6) Scott Alfeld, Jerry Zhu and Paul Barford. "Data Poisoning Attacks against Autoregressive Models", In Proceedings of the Conference on Artificial Intelligence (AAAI), Phoenix, AZ, February, 2016.

7) Aaron Cahn, Scot Alfeld, Paul Barford and S. Muthukrishnan. "An Empirical Study of Web Cookies", In Proceedings of the World Wide Web Conference (WWW '16), Montreal, CA, April, 2016.

8) Scott Alfeld, Jerry Zhu and Paul Barford. "Machine Teaching as Search", In the Symposium on Combinatorial Search - short paper, Tarrytown, NY, July, 2016

9) Ram Durairajan and Paul Barford. "A Techno-Economic Framework for Broadband Deployment in Underserved Areas", In Proceedings of the ACM SIGCOMM Global Access to

the Internet for All (GAIA) Workshop", Florianpolis, Brazil, August, 2016. (**Best Paper Award**)

10) Aaron Cahn, Scott Alfeld, Paul Barford and S. Muthukrishnan. "What's in the Community Cookie Jar?", In Proceedings of the IEEE/ACM Conference on Advances in Social Network Analysis and Mining (ASONAM '16), San Francisco, CA, August, 2016.

11) Meena Syamkumar, Ram Durairajan and Paul Barford. "Bigfoot: A Geo-based Visualization Methodology for Detecting BGP Threats", In Proceedings of the IEEE Symposium on Visualization for Cybersecurity (VizSec), Baltimore, MD, October, 2016.

12) Sathiya Mani, Ram Durairajan, Paul Barford and Joel Sommers. "MNTP: Enhancing Time Synchronization for Mobile Devices," In Proceedings of the ACM Internet Measurement Conference, Santa Monica, CA, November, 2016.

13) Scott Alfeld, Jerry Zhu and Paul Barford. "Explicit Defense Actions Against Test-Set Attacks", In Proceedings of the Conference on Artificial Intelligence (AAAI), San Francisco, CA, February, 2017.

14) Ram Durairajan and Paul Barford. "A Techno-Economic Framework for Broadband Deployment in Underserved Areas", In ACM SIGCOMM Computer Communications Review, 47 (2), May, 2017.

15) Ram Durairajan, Sathiya Mani, Paul Barford, Rob Nowak and Joel Sommers. "Timeweaver: Opportunistic One Way Delay Measurement via NTP", Under submission to ACM SIGMETRICS Conference, 2018.

16) Meena Syamkumar, Sathiya Mani, Ram Durairajan, Paul Barford and Joel Sommers. "Wrinkles in Time: Detecting Internet-wide Events via NTP", Under submission to ACM SIGMETRICS Conference, 2018.

LIST OF ACRONYMS

ACM	Association for Computing Machinery
BGP	Boarder Gateway Protocol
CAIDA	Center for Applied Internet Data Analysis
DNS	Domain Name System servers
DHS	Department of Homeland Security
FCC	Federal Communications Commission
GIS	Geographic Information System
IMPACT	Information Marketplace for Policy and Analysis of Cyber-risk & Trust
IP	Internet Protocol
ISP	Internet Service Provider
IXP	Internet Exchange Points
Lat	Latitude
Lon	Longitude
MNTP	Mobile Network Timing Protocol
NTP	Network Timing Protocol
PREDICT	Protected Repository for the Defense of Infrastructure Against Cyber Threats
ROW	Right Of Way
SNTP	Simple Network Timing Protocol
US	United States